Signature linear model

Arts Lucie, Verbytska Olena

Introduction

In machine learning we often want to express an output $Y \in \mathbb{R}$ by a finite number of predictors which can be seen as a vector of $\mathbb{R}^p$. This is the case, for example, when we wish to determine the price of our ski holiday according to the number of people who will be part of the trip, the period during which we will spend our holiday or the ski resort in which we wish to go.

However, sometimes we want to predict an output based on data that cannot be identified to a vector. Indeed, when we want to determine whether a person is sick or not according to physiological variables as a function of time, such as their blood sugar level, this cannot be expressed as a vector of $\mathbb{R}^p$. In this case, it is a function $X : [0,1] \to \mathbb{R}^d$. Similarly, in order to identify whether or not a species of marine animals is present in a place, by means of a sound recording of what is happening in that space, the usual case cannot be used.

To solve this problem a natural idea is to extend the classical linear model to the case where the input X is a function $X : [0,1] \to \mathbb{R}^d, d \geq 1$.

In this course, we will focus on the signature linear model. We will first see what the signature is and give some properties. Then we will define the model and give its properties. Finally, we will apply it to real data.

1 Signature: Definition and first properties

In order to follow the vocabulary of the usual signature framework, we will call the function $X : [0,1] \to \mathbb{R}^d$ a path. We assume that X is of bounded variation, which means that X has finite length.

Definition 1.1. Let $X : [0,1] \to \mathbb{R}^d, t \mapsto (X_t^1, \ldots, X_t^d)^\top$. The total variation of X is defined by

$$\|X\|_{TV} = \sup_{\mathcal{I}} \sum_{(t_0,\ldots,t_k)\in\mathcal{I}} \left\|X_{t_i} - X_{t_{i-1}}\right\|,$$

where the supremum is taken over all finite subdivisions of $[0,1]$, and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$.
The set of paths of bounded variation is then defined by

$$BV(\mathbb{R}^d) = \{X : [0,1] \to \mathbb{R}^d \mid \|X\|_{TV} < \infty\}.$$

Remark 1.2. $BV(\mathbb{R}^d)$ endowed with the norm $\|\cdot\|_{BV(\mathbb{R}^d)}$ is a Banach space, where

$$\|X\|_{BV(\mathbb{R}^d)} = \|X\|_{TV} + \sup_{t\in[0,1]} \|X_t\|.$$

Definition 1.3. Let $X \in BV(\mathbb{R}^d)$ and $I = (i_1, \dots, i_k) \subset \{1, \dots, d\}^k$, $k \geq 1$, be a multi-index of length k. The signature coefficient of X along the index I on $[0,1]$ is defined by

$$S^I(X) = \int\cdots\int_{0\leq u_1<\cdots<u_k\leq 1} dX^{i_1}_{u_1} \dots dX^{i_k}_{u_k}.$$

$S^I(X)$ is then said to be a signature coefficient of order k.
The signature of X is the sequence containing all signature coefficients :

$$S(X) = (1, S^{(1)}(X), \dots, S^{(d)}, S^{(1,1)}(X), S^{(1,2)}(X), \dots, S^{(i_1,\dots,i_k)}(X), \dots).$$

The signature of X truncated at order m, denoted by $S^m(X)$, is the sequence containing all signature coefficients of order lower than or equal to m, that is

$$S^m(X) = (1, S^{(1)}(X), S^{(2)}(X), \dots, S^{(d,\dots,d)}(X)),$$

with $(d, \dots, d)$ of length m.

Remark 1.4.

- The definition can be extended to path $X : [s,t] \to E$ by changing the bounds of the integrals.
- The signature is invariant by translation.
- There are d^k signature coefficients of order k. Therefore, the signature of X truncated at order m is a vector of dimension $s_d(m)$, where

$$d_d(m) = \sum_{k=0}^{m} d^k = \begin{cases} m+1, & d = 1; \\ \frac{d^{m+1}-1}{d-1}, & d \geq 2. \end{cases}$$

Example 1.5. Let X be a parameterized curve: for any $t \in [0,1]$, $X_t = (t, f(t))$, where $f : \mathbb{R} \to \mathbb{R}$ is a smooth function. Then,

$$S^{(1)}(X) = \int_0^1 dX^1_t = \int_0^1 dt = 1, \quad S^{(2)}(X) = \int_0^1 dX^2_t = \int_0^1 f'(t)dt = f(1) - f(0),$$

where f' denotes the derivative of f. Similarly, the signature coefficient along (1,2) is

$$S^{(1,2)}(X) = \int_0^1 \int_0^1 dX^1_u dX^2_t = \int_0^1 \left(\int_0^1 du\right) f'(t)dt = \int_0^1 tf'(t)dt = f(1) - \int_0^1 f(t)dt.$$

Example 1.6. Let X be a d-dimensional linear path:

$$X_t = \begin{pmatrix} X_t^1 \\ \vdots \\ X_t^d \end{pmatrix} = \begin{pmatrix} a_1 + b_1 t \\ \vdots \\ a_d + b_d t \end{pmatrix}$$

Then, for any index $I = (i_1, \dots, i_k) \subset \{1, \dots, d\}^k$, the signature coefficient along I is

$$S^{(i_1,\dots,i_k)}(X) = \int\limits_{0 \le u_1 < \cdots < u_k \le 1} \cdots \int dX_{u_1}^{i_1} \cdots dX_{u_k}^{i_k} = \int\limits_{0 \le u_1 < \cdots < u_k \le 1} \cdots \int b_{i_1} du_1 \cdots b_{i_k} du_k = \frac{b_{i_1} \cdots b_{i_k}}{k!}.$$

Definition 1.7. Let $X : [0, s] \to E$ and $Y : [s, t] \to E$ be two continuous paths.
Their concatenation is the path $X * Y$ defined by

$$(X * Y)_u = \begin{cases} X_u, & u \in [0, s]; \\ X_s + Y_u - Y_s, & u \in [s, t]. \end{cases}$$

Theorem 1.8 (Chen's Identity)**.** *Let $0 \le s \le t$, $X : [0, s] \to E$ and $Y : [s, t] \to E$ be two continuous paths. Then it holds that*

$$S(X * Y) = S(X) \otimes S(Y),$$

$$i.e.\ \ S^{(i_1,\dots,i_k)}(X * Y) = \sum_{\ell=0}^{k} S^{(i_1,\dots,i_\ell)}(X) \cdot S^{(i_{\ell+1},\dots,i_k)}(Y).$$

Proof. [2] Let $Z = X * Y : [s, u] \to E$ and $S(Z)$ its signature. We look at the order n of its signature:

$$\begin{aligned} S(Z)^n &= \int\limits_{s < u_1 < u_2 < \cdots < u_n < u} \cdots \int dZ_{u_1} \otimes \cdots \otimes dZ_{u_n} \\ &= \sum_{k=0}^{n} \int\limits_{s < u_1 < \cdots < u_k < t < u_{k+1} < \cdots < u_n < u} \cdots \int dZ_{u_1} \otimes \cdots \otimes dZ_{u_n} \\ &= \sum_{k=0}^{n} \int\limits_{s < u_1 < \cdots < u_k < t} \cdots \int dX_{u_1} \otimes \cdots \otimes dX_{u_k} \otimes \int\limits_{t < u_{k+1} < \cdots < u_n < u} \cdots \int dY_{u_{k+1}} \otimes \cdots \otimes dY_{u_n} \\ &= \sum_{k=0}^{n} X^k \otimes Y^{n-k}, \end{aligned}$$

where we used the Fubini's theorem. Therefore, $S(Z) = S(X) \otimes S(Y)$.

□

Proposition 1.9. *Assume that $X \in BV(\mathbb{R}^d)$ contains at least one monotone coordinate, then $S(X)$ characterizes X up to translations and reparameterizations.*

Remark 1.10. For any path $X \in BV(\mathbb{R}^d)$, the time-augmented path $\widetilde{X_t} = (X_t, t)^\top \in BV(\mathbb{R}^{d+1})$ satisfies the assumption of Proposition 1.9, which ensure signature uniqueness. That's why we will always use this time-augmentation transformation before computing signatures.

Theorem 1.11. *Suppose $f : E_1 \to \mathbb{R}$ is a continuous function where E_1 is a compact subset of E. Then for every $\varepsilon > 0$, there exists a linear functional $L \in E^{\otimes n}$ such that for every $a \in E_1$,*

$$|f(a) - L(a)| \leq \varepsilon.$$

The proof can be found in [6], theorem 3.1.

Corollary 1.12. *Let $D \subset BV(\mathbb{R}^d)$ be a compact set of paths such that, for any $X \in D$, $X_0 = 0$, and denote by $\widetilde{X} = (X_t, t)_{t\in[0,1]}^\top$ the associated time-augmented path. Let $f : D \to \mathbb{R}$ be a continuous function. Then, for every $\varepsilon > 0$, there exist $m^* \in \mathbb{N}, \beta^* \in \mathbb{R}^{s_d(m^*)}$, such that, for any $X \in D$,*

$$|f(X) - \langle \beta^*, S^{m^*}(\widetilde{X})\rangle| \leq \varepsilon,$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product on $\mathbb{R}^{s_d(m^*)}$.

Remark 1.13. [9] This result shows that any continuous function on a compact set of paths can be approximated arbitrarily well simply by a linear combination of terms of the signature.

Proposition 1.14. *Let $X : [0, 1] \to \mathbb{R}^d$ be a path in $BV(\mathbb{R}^d)$. Then for any $m \geq 0$,*

$$\|S^m(X)\| \leq \sum_{k=0}^{m} \frac{\|X\|_{TV}^k}{k!} \leq e^{\|X\|_{TV}}.$$

2 The signature linear model

We now have all the definitions and properties we need to define the signature linear model.

2.1 The model

Let's assume that we have a data set $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ where the pairs (X_i, Y_i) are independent and identically distributed copies of a random couple (X, Y), where X is a function, $X \in BV(\mathbb{R}^d)$, and $Y \in \mathbb{R}$. We want to find a relationship between Y and X. The goal is to approximate the regression function $f(X) = \mathbb{E}[Y|X]$.

Assumption 2.1. We assume that $d \geq 2$ and X is a time-augmented path.

Definition 2.2 (first version of the model)**.** [6] Let X and Y be two stochastic processes taking values in E and W respectively. Suppose that the signatures of X and Y denoted by S_X and S_Y are well defined a.s. Assume that

$$S_Y = L(S_X) + \varepsilon,$$

where $\mathbb{E}[\varepsilon|S_X] = 0$ and L is a linear functional mapping $E^{\otimes n}$ to $W^{\otimes n}$.

Definition 2.3 (more recent version of the model)**.** Let $Y \in \mathbb{R}$ and $X \in BV(\mathbb{R}^d)$. We assume that there exists $m \in \mathbb{N}$, $\beta_m^* \in \mathbb{R}^{s_d(m)}$, such that

$$\mathbb{E}[Y|X] = \langle \beta_m^*, S^m(X) \rangle, \quad Var(Y|X) \leq \sigma^2 < \infty.$$

Remark 2.4. In this course, we consider the smallest $m = m^* \in \mathbb{N}$ such that there exists $\beta_{m^*}^* \in \mathbb{R}^{s_d(m^*)}$ satisfying $\mathbb{E}[Y|X] = \langle \beta_{m^*}^*, S^{m^*}(X) \rangle$. It is a regression model where the regression function is a linear form on the signature.

Remark 2.5. It is a very general model because the hypotheses are very weak: $\mathbb{E}[Y|X]$ must be continuous and $S(X)$ must characterize X.

Proposition 2.6. *Under assumption that the data is in a compact set, for any $\varepsilon > 0$, there exists $m^* \in \mathbb{N}$ and $\beta_{m^*}^* \in \mathbb{R}^{s_d(m^*)}$ such that*

$$|\mathbb{E}[Y|X] - \langle \beta_{m^*}^*, S^{m^*}(X) \rangle| \leq \varepsilon.$$

Proof. It is a direct application of the corollary 1.12. □

Remark 2.7. There are two unknown quantities in the model:

- m^* which is the truncation order of the signature of X;
- $\beta_{m^*}^*$ which is the vector of the regression coefficients and is directly related to m^*.

2.2 The estimation of m^*

Definition 2.8. The ball in $\mathbb{R}^{s_d(m)}$ of radius α centered at 0 is denoted by

$$B_{m,\alpha} = \left\{ \beta \in \mathbb{R}^{s_d(m)} \mid \|\beta\| \leq \alpha \right\}.$$

Definition 2.9. The theoretical risk is defined by $\mathcal{R}_m(\beta) = \mathbb{E}[Y - \langle \beta, S^m(X) \rangle]^2$.

Notation 2.10. $L(m) = \inf_{\beta \in B_{m,\alpha}} \mathcal{R}_m(\beta) = \mathcal{R}_m\left(\beta_m^*\right)$.

Remark 2.11. Note that β_m^* exists because the problem is convex.

Remark 2.12. $L(m) = \mathbb{E}[Var(Y|X)]$.

Assumption 2.13. (H_α) We assume that there exists $\alpha > 0$ such that $\beta^*_{m^*} \in B_{m^*,\alpha}$.

Definition 2.14. The empirical risk with signature truncated at order m is

$$\widehat{\mathcal{R}}_{m,n}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \beta, S^m(X_i)\rangle)^2.$$

And it's minimum is

$$\widehat{L}_n(m) = \min_{\beta \in B_{m,\alpha}} \widehat{\mathcal{R}}_{m,n}(\beta) = \widehat{\mathcal{R}}_{m,n}(\widehat{\beta}_m).$$

Definition 2.15. An estimator of m^* is

$$\widehat{m} = \min\left(\operatorname*{argmin}_{m\in\mathbb{N}}\left(\widehat{L}_n(m) + \mathrm{pen}_n(m)\right)\right),$$

where $m \to pen_n(m)$ is an increasing function of m.

3 Model performance

Assumption 3.1. (H_K) There exists $K_Y > 0$ and $K_X > 0$ such that a.s. $|Y| \leq K_Y$ and $\|X\|_{TV} \leq K_X$.

Notation 3.2. $K = 2\left(K_Y + \alpha e^{K_Y}\right)e^{K_X}$.

Theorem 3.3. *Let* $K_{\mathrm{pen}} > 0,\ 0 < \rho < \frac{1}{2}$, *and*

$$\mathrm{pen}_n(m) = K_{\mathrm{pen}} n^{-\rho}\sqrt{s_d(m).}$$

Let n_0 *be the smallest integer satisfying*

$$(n_0)^{\tilde{\rho}} \geq (432K\alpha\sqrt{\pi} + K_{\mathrm{pen}})\left(\frac{2\sqrt{s_d(m^*+1)}}{L(m^*-1)-\sigma^2} + \frac{\sqrt{2s_d(m^*+1)}}{K_{\mathrm{pen}}\sqrt{d^{m^*+1}}}\right),$$

where $\tilde{\rho} = \min(\rho, \frac{1}{2} - \rho)$. *Then, under the assumptions* (H_α) *and* (H_K) *[2.13, 3.1], for any* $n \geq n_0$,

$$\mathbb{P}(\widehat{m} \neq m^*) \leq C_1 \exp\left(-C_2 n^{1-2\rho}\right),$$

where the constants C_1 *and* C_2 *are defined by*

$$C_1 = 74\sum_{m>0} e^{-C_3 s_d(m)} + 148m^*, \quad C_3 = \frac{K^2_{\mathrm{pen}} d^{m^*+1}}{128 s_d(m^*+1)(72K^2\alpha^2 + K_Y^2)},$$

and

$$C_2 = \frac{1}{16\left(1152K^2\alpha^2 + K_Y^2\right)}\min\left(\frac{K^2_{pen}\, d^{m^*+1}}{8s_d\left(m^*+1\right)}, L(m^*-1) - \sigma^2\right).$$

Proof. The detailed proof of this theorem can be found in [3], Section 8. It is long and based on chaining tail inequalities that bound uniformly the tails of the risk.

□

Now let us discuss the behavior of the constants when the different parameters vary:

- d increases $\implies d^{m^*+1} \sim s_d\,(m^*+1)$ and C_1, C_2 stay of the same order (provided that the risk $L\,(m^*-1) \equiv const$).
  Therefore, the quality of the bound does not change in high dimensions.
  $n_0 = O(d^{m^*/2\tilde{\rho}})$: we need exponentially more data when d grows.

- m^* is large $\implies$ C_2 and C_3 stay of the same order, $C_1 \sim 148m^*$, $n_0 = O\left(d^{m^*/2\tilde{\rho}}\right)$.
  The size of the coefficient $\beta^*_{m^*}$ increases and therefore more data are needed to estimate it.

- α increases $\implies$ n_0 and C_1 increase, C_2 decreases.
  $B_{m,\alpha}$ gets larger for any m, therefore, more data is needed and the quality of the estimator deteriorates.

- $(L(m^*-1)-\sigma^2) > 0$ is close to 0 $\implies$ a model truncated at m^*-1 is almost as good as a model truncated at m^*, since $L\,(m^*-1)-\sigma^2 \le L\,(m^*-1)-L\,(m^*)$.
  This difference decreases $\implies$ n_0 increases, C_2 decreases.
  When it is harder to find that a truncation order of m^* is better than m^*-1, then the estimator $\widehat{m}$ deteriorates.

Remark 3.4. This theorem provides a non-asymptotic bound on the convergence of $\widehat{m}$. It implies the almost sure convergence of $\widehat{m}$ to m^*. The penalty includes an arbitrary constant K_{pen} . Its value that minimizes n_0 is

$$K^*_{\text{pen}} = \sqrt{\frac{(L\,(m^*-1)-\sigma^2)\,432\sqrt{\pi}\alpha K}{d^{m^*+1}}}.$$

Having $\widehat{m}$ one can choose to estimate $\beta^*_{m^*}$ by $\hat{\beta}_{\widehat{m}}$. Then we can get the following bound.

Corollary 3.5. *Under the assumptions (H_α) and (H_K), for any $n \ge n_0$,*

$$\mathbb{E}\left(\left\langle \hat{\beta}_{\widehat{m}}, S^{\widehat{m}}(X)\right\rangle - \left\langle \beta^*_{m^*}, S^{m^*}(X)\right\rangle\right)^2 \le \frac{C_5}{\sqrt{n}} + C_6 e^{-C_2 n^{1-2\rho}},$$

where the constants C_5 and C_6 are defined by

$$C_5 = 36K\alpha\sqrt{\pi}\,(m^*+1)\,\sqrt{s_d\,(m^*)}, \quad C_6 = 2664K\alpha\sqrt{\pi}\sum_{m>m^*}\sqrt{s_d(m)}e^{-C_3 s_d(m)} + 2\alpha^2 e^{K_X}C_1.$$

To prove the corollary we require the following lemma, proof of which can be found in [3], Section 8, Lemma 4.

Lemma 3.6. *Let $K_{pen} > 0$, $0 < \rho < \frac{1}{2}$, and $\mathrm{pen}_n(m) = K_{\mathrm{pen}} n^{-\rho}\sqrt{s_d(m)}$. Then*

$$\mathbb{E}\left[\sup_{\beta\in B_{\bar{m},\alpha}}\left|\widehat{\mathcal{R}}_{\widehat{m},n}(\beta) - \mathcal{R}_{\widehat{m}}(\beta)\right|\right] \leq 36K\alpha\sqrt{\frac{\pi}{n}}\left((m^*+1)\sqrt{s_d(m^*)} + 74e^{-C_3 n^{1-2\rho}}\sum_{m>m^*}\sqrt{s_d(m)}e^{-C_3 s_d(m)}\right).$$

Proof of Corollary 3.5.

1. Let us note that $\mathbb{E}\left(\left\langle\widehat{\beta}_{\widehat{m}}, S^{\widehat{m}}(X)\right\rangle - \left\langle\beta^*_{m^*}, S^{m^*}(X)\right\rangle\right)^2 = \mathbb{E}\left(\mathcal{R}_{\widehat{m}}\left(\widehat{\beta}_{\widehat{m}}\right)\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)$. Moreover, we have a.s.

$$\begin{aligned}
\mathcal{R}_{\widehat{m}}\left(\widehat{\beta}_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right) &= \mathcal{R}_{\widehat{m}}\left(\widehat{\beta}_{\widehat{m}}\right) - \mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) + \mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)\\
&= \mathcal{R}_{\widehat{m}}\left(\widehat{\beta}_{\widehat{m}}\right) - \widehat{\mathcal{R}}_{\widehat{m},n}\left(\widehat{\beta}_{\widehat{m}}\right) + \underbrace{\widehat{\mathcal{R}}_{\widehat{m},n}\left(\widehat{\beta}_{\widehat{m}}\right) - \widehat{\mathcal{R}}_{\widehat{m},n}\left(\beta^*_{\widehat{m}}\right)}_{=\min_{\beta\in B_{\widehat{m},\alpha}}\widehat{\mathcal{R}}_{\widehat{m},n}(\beta) - \widehat{\mathcal{R}}_{\widehat{m},n}\left(\beta^*_{\widehat{m}}\right)\leq 0}\\
&\quad + \widehat{\mathcal{R}}_{\widehat{m},n}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) + \mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)\\
&\leq \mathcal{R}_{\widehat{m}}\left(\widehat{\beta}_{\widehat{m}}\right) - \widehat{\mathcal{R}}_{\widehat{m},n}\left(\widehat{\beta}_{\widehat{m}}\right) + \widehat{\mathcal{R}}_{\widehat{m},n}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) + \mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)\\
&\leq 2\sup_{\beta\in B_{\widehat{m},\alpha}}\left|\widehat{\mathcal{R}}_{\widehat{m},n}(\beta) - \mathcal{R}_{\widehat{m}}(\beta)\right| + \mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)
\end{aligned}$$

2. Let us now prove the following equality

$$\mathbb{E}\left(\mathcal{R}_{\bar{m}}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)\right) = 2\alpha^2 e^{K_X} C_1 e^{-C_2 n^{1-2\rho}}.$$

Since, for any $m \in \mathbb{N}$, $\left\langle\beta^*_m, S^m(X)\right\rangle^2 \leq \left\|\beta^*_k\right\|_2^2 \left\|S^m(X)\right\|_2^2 \leq \alpha^2 e^{K_X}$, it follows that

$$\begin{aligned}
\mathbb{E}\left(\mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)\right) &= \mathbb{E}\left(\left(Y - \left\langle\beta^*_{\widehat{m}}, S^{\widehat{m}}(X)\right\rangle\right)^2 - \left(Y - \left\langle\beta^*_{m^*}, S^{m^*}(X)\right\rangle\right)^2\right)\\
&= \mathbb{E}\left(\left(\left\langle\beta^*_{m^*}, S^{m^*}(X)\right\rangle + \varepsilon - \left\langle\beta^*_{\widehat{m}}, S^{\widehat{m}}(X)\right\rangle\right)^2 - \varepsilon^2\right)\\
&= \mathbb{E}\left(\left(\left\langle\beta^*_{m^*}, S^{m^*}(X)\right\rangle - \left\langle\beta^*_{\widehat{m}}, S^{\widehat{m}}(X)\right\rangle\right)^2\right) \leq 2\alpha^2 e^{K_X}\mathbb{P}\left(\widehat{m} \neq m^*\right)
\end{aligned}$$

By Theorem 3.3, this yields $\mathbb{E}\left(\mathcal{R}_{\widehat{m}}\left(\beta^*_{\widehat{m}}\right) - \mathcal{R}_{m^*}\left(\beta^*_{m^*}\right)\right) \leq 2\alpha^2 e^{K_X} C_1 e^{-C_2 n^{1-2\rho}}$.

Letting $C_5 = 36K\alpha\sqrt{\pi}\left(m^*+1\right)\sqrt{s_d\left(m^*\right)}$, and $C_6 = 2664K\alpha\sqrt{\pi}\sum_{m>m^*}\sqrt{s_d(m)}e^{-C_3 s_d(m)} + 2\alpha^2 e^{K_X} C_1$, since, by intermediate result of proof of Theorem 3.3, $C_2 \leq C_3$, and using Lemma 3.6, we conclude that

$$\mathbb{E}\left(\langle\widehat{\beta}_{\widehat{m}}, S^{\widehat{m}}(X)\rangle - \langle\beta^*_{m^*}, S^{m^*}(X)\rangle\right)^2 \leq \frac{C_5}{\sqrt{n}} + C_6 e^{-C_2 n^{1-2\rho}}.$$

□

4 Computational aspects

4.1 The signature linear model algorithm

To be able to compute signatures efficiently, we can choose to interpolate the sampled points linearly. This allows to reduce our problem to computing signatures of piecewise linear paths. Definition 1.7 of concatenation and Chen's theorem 1.8 provide a formula to compute the signature of a concatenation of two paths.

To compute the signature of a piecewise linear path, it is sufficient to iterate the following two steps:

- Compute the signature of a linear section of the path like we did in Example 1.6.
- Concatenate it to the other pieces with Chen's theorem.

This procedure is implemented in the Python library *iisignature* [7]. Thus, for a sample consisting of p points in $\mathbb{R}^d$, if we consider the path formed by their linear interpolation, the computation of the path signature truncated at level m takes $O(pd^m)$ operations. The complexity is therefore linear in the number of sampled points but exponential in the truncation order m, which emphasizes once more the importance of the choice of $\widehat{m}$.

In practice, we are given a dataset $\{(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)\}$, where, for any $1 \leq i \leq n, Y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^{d \times p_i}$. The columns of the matrix $\mathbf{x}_i$ correspond to values of a process X_i in $\mathbb{R}^d$ sampled at p_i different times. We fix $M \in \mathbb{N}$ such that, for any $m \geq M$, the function $m \mapsto \widehat{L}_n(m) + \mathrm{pen}_n(m)$ is strictly increasing and apply the procedure described in Algorithm 1. The parameter ρ is set to 0.4. The constant K_{pen} is calibrated with the so-called slope heuristics method [1].

Algorithm 1: Pseudo-code for the signature linear model.

Data: $\{(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)\}$.
Result: Estimators $\widehat{m}$ and $\widehat{\beta}_{\widehat{m}}$.

- Interpolate linearly the columns of $\mathbf{x}_i$ so as to have a set of continuous piecewise linear parts $X_i : [0,1] \to \mathbb{R}^d, 1 \leq i \leq n$. Add a time dimension i.e. consider the path $\widetilde{X}_i : [0,1] \to \mathbb{R}^{d+1}$, where $\widetilde{X}_i^j = X_i^j$ for $1 \leq j \leq d$, and $X_{i,t}^{d+1} = t, t \in [0,1]$.
- Select the Ridge regularization parameter λ by cross validation on the regression model with $\{S^1(\widetilde{X}_1) \dots, S^1(\widetilde{X}_n)\}$ as predictors.

for $m = 1, \cdots, M$ **do**

- Compute signatures truncated at level $m : \{S^m(\widetilde{X}_1), \dots, S^m(\widetilde{X}_n)\}$.
- Fit a Ridge regression on the pairs $\{(S^m(\widetilde{X}_1), Y_1), \dots, (S^m(\widetilde{X}_n), Y_n)\}$.
- Compute its squared loss $\widehat{L}_n(m)$.
- Compute the penalization $\mathrm{pen}_n(m) \leftarrow K_{\mathrm{pen}} \frac{\sqrt{s_d(m)}}{n^\rho}$.

end

- Choose $\widehat{m} \leftarrow \underset{0 \leq m \leq M}{\mathrm{argmin}} (\widehat{L}_n(m) + \mathrm{pen}_n(m))$.
- Compute $\widehat{\beta}_{\widehat{m}}$ by fitting a Ridge regression on $\{(S^{\widehat{m}}(\widetilde{X}_1), Y_1), \dots, (S^{\widehat{m}}(\widetilde{X}_n), Y_n)\}$: $\widehat{\beta}_{\widehat{m}} \leftarrow (\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{I})^{-1} \mathbf{S}^\top \mathbf{Y}$, where $\mathbf{S} \in \mathbb{R}^{n \times s_d(\widehat{m})}$ is the matrix which rows are the signatures of the inputs $S^{\widehat{m}}(\widetilde{X}_i)^\top, \mathbf{I} \in \mathbb{R}^{s_d(\widehat{m}) \times s_d(\widehat{m})}$ is the identity matrix, and $\mathbf{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^d$ is the vector of responses.

4.2 Application to EEG data

In this section we adapted the code [5] to the "EEG Eye State" dataset [8]. The data is from one continuous EEG measurement with the Emotiv EEG Neuroheadset. The eye state was detected via a camera during the EEG measurement and added later manually to the file after analysing the video frames. **'1'** indicates the eye-closed and **'0'** the eye-open state.

We study if the eyes are open using the data from 3 channels: **F7, FC5, FC6**. We consider two situations for the predictor function X: a univariate and a multivariate case. In the univariate case, we are given the values of the channel **F7**. In this case, the data is in dimension $d = 1$ and sampled at $p = 168$ values. In the multivariate case, we add the information from channels **FC5** and **FC6** to X, making it a path in dimension $d = 3$. We show in Figure 1 one sample in the multivariate case (in the univariate case, X consists only of the blue solid curve).

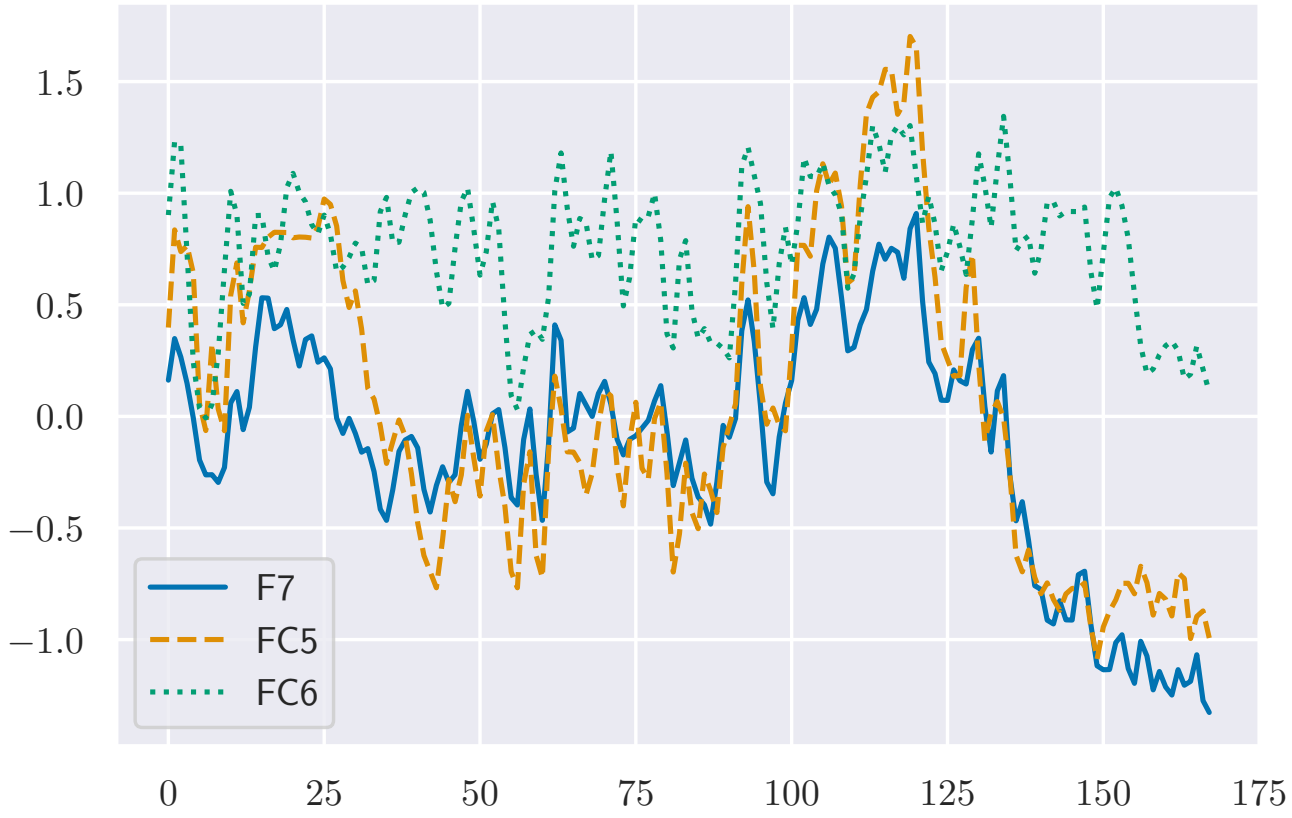
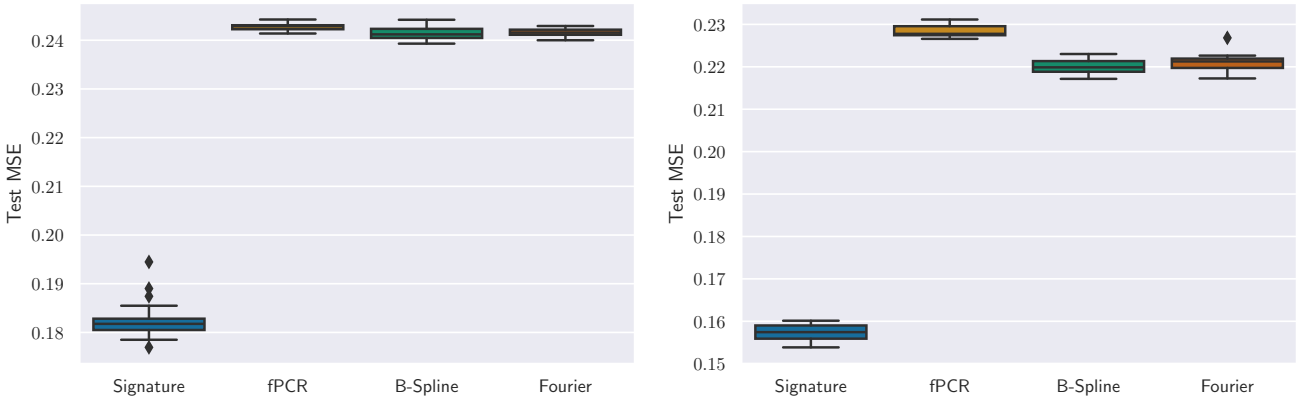


Figure 1: One sample from the "EEG Eye State" dataset

We demonstrate performance of signature linear model compared to canonical approaches in functional data analysis for our data. We compare our model to functional principal component regression (fPCR), and to two functional linear models (we consider two choices for the basis namely the B-Spline and Fourier basis).

The approach consists in projecting the function $X : [0, 1] \to \mathbb{R}^d$ onto the basis functions, coordinate by coordinate. The number of basis functions is selected via cross-validation (with a minimum of 4 and maximum of 14 for Fourier and B-Splines, and a minimum of 1 and a maximum of 6 for the fPCR). For the fPCR, we first smooth the functional covariates with 7 B-Splines. This procedure is implemented with the Python package *scikit-fda*.



(a) Univariate case

(b) Multivariate case

Figure 2: Test MSE for different regression models for the "EEG Eye State" dataset

We perform 20 random train/test splits and show in Figure 2 a boxplot of the test MSE for each model. In the both cases for our dataset the signature linear model performs significantly better than other. As we can see the range of MSE for signature linear model decreases while the dimension increases.

For another dataset it is shown in [3], that in the univariate case, the B-Splines linear model performs the best. However, in the multivariate case the signature model's error has halved while the others have not changed much.

To see the illustrations of the different steps of Algorithm 1 and the convergence of the estimator $\widehat{m}$ with simulated data visit Section 6 in [3].

5 Conclusion

The signature is a useful tool for summarizing multidimensional time series data, applicable in various situations. This outline is a first step in understanding how to use it in statistics, with many potential applications. Here we focused on linear regression, but signatures are also relevant in classification or unsupervised learning. The main challenge is the high dimension of the regression coefficient, which depends exponentially on m. This time it was handled by carefully choosing the truncation order. However, this is not the only option. For example, regularization approaches that induce a sparsity pattern on this coefficient, or the use of a related lower dimensional object called the logsignature, can be applied.

As a continuation of this course we recommend to get acquainted with the embedding and learning with signatures (for example, [4]).

References

[1] L. Birge and P. Massart. "Minimal penalties for gaussian model selection". In: *Probability Theory and Related Fields* 138 (2007), pp. 33–73.

[2] A. Fermanian. "Signature and statistical learning". In: (2018). URL: https://afermanian.github.io/assets/docs/master_thesis_fermanian.pdf.

[3] A. Fermanian. "Linear functional regression with truncated signatures". In: *arXiv:2006.08442* (2020).

[4] A. Fermanian. "Embedding and learning with signatures". In: *Computational Statistics and Data* 157 (2021), pp. 107–148. URL: https://hal.science/hal-02387258.

[5] A. Fermanian. *Functional linear regression with truncated signatures*. GitHub Repository. 2021. URL: https://github.com/afermanian/signature-regression/blob/master/README.md.

[6] D. Levin, T. Lyons, and H. Ni. "Learning from the past, predicting the statistics for the future, learning an evolving system". In: *arXiv:1309.0260v6* (2013).

[7] J. Reizenstein and B. Graham. "Algorithm 1004: The iisignature library: Efficient calculation of iterated-integral signatures and log signatures". In: *ACM Transactions on Mathematical Software* (2020).

[8] O. Roesler. *EEG Eye State*. UCI Machine Learning Repository. 2013. DOI: https://doi.org/10.24432/C57G7J.

[9] M. Wiese, P. Murray, and R. Korn. "Sig-Splines: universal approximation and convex calibration of time series generative models". In: *arXiv:2307.09767v1* (2023).